AbstractID: 9731 Title: Machine Learning based Clinical Research: The example of Lung Cancer

Machine Learning based Clinical Research: The example of Lung Cancer

P. Lambin[1], C. Dehing-Oberije[1], L. Persoon[1], S. Yu[2], D. De Ruysscher[1], R. B. Rao[2], A. Dekker[1].

[1] University Medical Center Maastricht, Department of Radiation Oncology, MAASTRO clinic, GROW, The Netherlands

[2] Siemens Medical Solutions USA, Inc., Malvern, Pennsylvania, USA

Maximum 500 words without titels authors see: Invited Speaker - Abstract/Handout Preparation Instructions to Authors:

http://www.aapm.org/meetings/08AM/invitedhelp.asp

Abstract is due: Friday, April 11, 5PM EST

Purpose: The *hypothesis* of this long term project is that a multicentric based information system based on four modules (multiparametric interconnected healthcare databases, data mining tools, updated machine learning based predictive algorithms and user interfaces) will facilitate and accelerate research in oncology. We call this approach “Machine Learning Based Clinical Research (MLBCR)”. We performed a pilot project in non-small cell lung cancer (NSCLC) patients for which clinical TNM stage is highly inaccurate for the prediction of survival of non-surgical patients and alternatives are currently lacking. The objectives of this study were to develop and validate a prediction model for survival of NSCLC patients, treated with (chemo) radiotherapy, using clinical factors.

Patients and Methods: Three interconnected databases were mirrored into a data warehouse using a disease based , cohort-specific data model. The three data sources were a) electronic medical records, b) imaging and DICOM-RT objects in a RT-PACS and c) treatment information in a record and verify database. Data from 403 consecutive inoperable NSCLC patients, stage I-IIIB, treated radically with (chemo) radiation were selected. In 82 patients data from blood samples were available. The 2-norm Support Vector Machines were used to build the prognostic models. Performance of the models was expressed as the AUC (Area Under the Curve) of the Receiver Operating Characteristic (ROC) and assessed using leave-one-out (LOO) cross-validation. The prognostic model, using clinical factors only, was validated using two external, independent datasets with 36 and 65 patients, respectively. In addition, a risk score was calculated and a nomogram, which is in fact a graphical representation of the risk score, was made for practical use.

Results: The model, based on 403 patients and using clinical factors, consisted of gender, WHO performance status, forced expiratory volume (FEV_1), number of positive lymph node stations on PET and gross tumor volume (on PET-CT). The AUC, assessed by LOO cross-validation, was 0.75 (95% CI 0.70-0.82), while application of the model to the external datasets yielded an AUC of 0.75 and 0.76 respectively. Splitting the MAASTRO cohort into 3 subgroups, based on the risk score, resulted in the identification of a high, medium and low risk group. The 2-year survival was 66% (95% CI 54%-78%) for the low risk group, 29% (95% CI 21%-37%) for the medium risk group and 14% (95% CI 5%-23%) for the high risk group. If blood biomarkers were available, based on the 82 patients the prognostic model consisted of three additional biomarkers factors: OPN, IL8 and CEA. The LOO AUC was 0.83 (95% CI 0.76-0.94), which is significantly better than the prognostic model using only clinical factors based on the same 82 patients (AUC 0.71, 95% CI 0.60-0.87). ***In conclusion***, the model, using clinical factors, successfully estimates 2-year survival of NSCLC patients and the performance, assessed internally as well as in two independent datasets, is good. Combining blood biomarkers with clinical factors yielded a significantly better performance than using clinical factors only (AUC: 0.83 vs 0.71). We concluded that MLBCR is feasible. The bottle neck is the availability of external data sets. Therefore, we need to invest in international standards as well in multicentric approaches allowing to recruit more patients, preferably having had different type of treatments, and to have quick access to external validation data sets.

Conflict of interest: This project has been partially funded by Siemens IKM.